Leverage in penalized least squares estimation

Keith Knight
University of Toronto
keith@utstat.toronto.edu

Abstract

The German mathematician Carl Gustav Jacobi showed in 1841 that least squares estimates can be written as an expected value of elemental estimates with respect to a probability measure on subsets of the predictors; this result is trivial to extend to penalized least squares where the penalty is quadratic. This representation is used to define the leverage of a subset of observations in terms of this probability measure. This definition can be applied to define the influence of the penalty term on the estimates. Applications to the Hodrick-Prescott filter, backfitting, and ridge regression as well as extensions to non-quadratic penalties are considered.

1 Introduction

We consider estimation in the linear regression model

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n)$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters.

Given observations $\{(\boldsymbol{x}_i, y_i) : i = 1, \cdots, n\}$, we define the penalized least squares (PLS) estimate $\widehat{\boldsymbol{\beta}}$ as the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \boldsymbol{\beta}^T A\boldsymbol{\beta}. \tag{1}$$

where A is a non-negative definite matrix with $A = L^TL$ for some $r \times p$ matrix L. Note that L is in general not uniquely specified since for any $r \times r$ orthogonal matrix O, $A = (OL)^T(OL)$. In the context of this paper, we would typically choose L so that its rows are interpretable in some sense.

Examples of PLS estimates include ridge estimates (Hoerl and Kennard, 1970), smoothing splines (Craven and Wahba, 1978), the Hodrick-Prescott filter (Hodrick and Prescott, 1997), and (trivially when $A = 0$) ordinary least squares (OLS) estimates. These estimates are useful, for example, the OLS is unstable or when $p > n$; if A has rank p then (1) will always

have a unique minimizer. Note that our definition of PLS includes only quadratic penalties and thus excludes methods with non-quadratic penalties such as the LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), among others. Extensions of the theory developed in this paper to the LASSO will be discussed briefly in section 8.

For OLS estimation, the leverage is typically defined in terms of the diagonal elements $h_{11}, \cdots, h_{nn}$ of the so-called "hat" matrix (Hoaglin and Welsch, 1978), which is an orthogonal projection onto the column space of the matrix whose rows are $\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T$. If $\widehat{y}_i = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{ols}$ is the fitted value based on the OLS estimate then

$$h_{ii} = \frac{\partial \widehat{y}_i}{\partial y_i}$$

and so the leverage can be thought of as the sensitivity of a given fitted value to a change in its observed response value. (More generally, we can define the equivalent degrees of freedom of a method as the sum over n observations of these partial derivatives.) This particular definition of leverage does not really work for the rows of L since the corresponding response values are equal to the fixed number 0 and so defining leverage as a partial derivative is not particularly sensible. Nonetheless, we would like to be able to extend the notion of leverage to a row (or rows) of L.

In "big data" problems where exact computation of OLS estimates is problematic, estimates based on subsamples of $\{(\boldsymbol{x}_i, y_i)\}$ are used where the observations are sampled with probabilities proportional to their leverages (or some estimates of these leverages); the idea is that biasing the sampling towards observations with higher leverage will result in a better sample for the purposes of approximating OLS estimates. More details on this method (which is called "algorithmic leveraging") can be found in Drineas *et al.* (2011, 2012), Ma *et al.* (2015), and Gao (2016). An extension of algorithmic leveraging called leveraged volume sampling is discussed in Derezinski *at al.* (2018).

Defining leverage for a set of observations is more ambiguous but very important; the leverage values $h_{11}, \cdots, h_{nn}$ often do not tell the whole story about the potential influence of a given observation. Cook and Weisberg (1982) propose using the maximum eigenvalue of the sub-matrix of the hat matrix whose rows and columns correspond to the indices of the observations. Clerc Bérod and Morgenthaler (1997) provide a geometric interpretation of this maximum eigenvalue. Similar influence measures for subsets of observations have been proposed; see Chatterjee and Hadi (1986) as well as Nurunnabi *et al.* (2014) for surveys of some of these methods. Hellton *et al.* (2019) study the influence of single observations in ridge regression.

The goal of this paper is to define the leverage of a set of observations in terms of a probability that reflects the influence that the set has on the estimate; in the case of single observations, this definition agrees with the standard definition of leverage. This definition will also allow us to define the leverage (or influence) of the penalty $\boldsymbol{\beta}^T A \boldsymbol{\beta}$ in (1) as well as the influence of individual rows or sets of rows of the matrix L defining A. In section

2, we will define the probability measure on subsets of size p that defines $\widehat{\boldsymbol{\beta}}$ minimizing (1) and describe its lower dimensional distributions. Our definition of leverage will be given in section 3 while sections 4 through 8 will discuss various applications.

2 Elemental estimation and PLS

The PLS estimates can be expressed as $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ where X and $\boldsymbol{y}$ are defined by

$$X = \begin{pmatrix} \mathcal{X} \\ L \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \\ \boldsymbol{x}_{n+1}^T \\ \vdots \\ \boldsymbol{x}_{n+r}^T \end{pmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ y_{n+1} \\ \vdots \\ y_{n+r} \end{pmatrix} \tag{2}$$

so that $\boldsymbol{x}_{n+1}^T, \cdots, \boldsymbol{x}_{n+r}^T$ are defined to be the rows of L and $y_{n+1} = \cdots = y_{n+r} = 0$. Note that $X^T X = \mathcal{X}^T \mathcal{X} + A$ and $X^T \boldsymbol{y} = \mathcal{X}^T \boldsymbol{y}_0$ where $\boldsymbol{y}_0 = (y_1, \cdots, y_n)^T$. We can also define the projection matrix onto the column space of X:

$$H = X(X^T X)^{-1} X^T = \begin{pmatrix} \mathcal{X}(\mathcal{X}^T \mathcal{X} + A)^{-1} \mathcal{X}^T & \mathcal{X}(\mathcal{X}^T \mathcal{X} + A)^{-1} L^T \\ L(\mathcal{X}^T \mathcal{X} + A)^{-1} \mathcal{X}^T & L(\mathcal{X}^T \mathcal{X} + A)^{-1} L^T \end{pmatrix}. \tag{3}$$

Jacobi (1841) showed that the OLS estimate can be written as a weighted sum of so-called elemental estimates and his result has been rediscovered several times in the interim, most notably by Subrahmanyam (1972); this result extends trivially to PLS estimates using elemental estimates based on the rows of X. (Berman (1988) extends Jacobi's result to generalized least squares.) In particular, if $s = \{i_1, \cdots, i_p\}$ is a subset of $\{1, \cdots, n+r\}$ then we can define the elemental estimate $\widehat{\boldsymbol{\beta}}_s$ satisfying

$$\boldsymbol{x}_{i_j}^T \widehat{\boldsymbol{\beta}}_s = y_{i_j} \quad \text{for } j = 1, \cdots, p$$

provided that the solution $\widehat{\boldsymbol{\beta}}_s$ exists. We can then write the PLS estimate as follows:

$$\widehat{\boldsymbol{\beta}} = \sum_s \frac{|X_s|^2}{\sum_u |X_u|^2} \widehat{\boldsymbol{\beta}}_s$$

where the summations are over all subsets of size p, $|K|$ denotes the determinant of a square matrix K and

$$X_s = \begin{pmatrix} \boldsymbol{x}_{i_1}^T \\ \boldsymbol{x}_{i_2}^T \\ \vdots \\ \boldsymbol{x}_{i_p}^T \end{pmatrix} \tag{4}$$

is a $p \times p$ sub-matrix of X whose row indices are in s. Therefore, we can think of the PLS estimate $\widehat{\boldsymbol{\beta}}$ as an expectation of elemental estimates with respect to a particular probability distribution; that is,

$$\widehat{\boldsymbol{\beta}} = E_{\mathcal{P}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}})$$

where the random subset $\mathcal{S}$ has a probability distribution

$$\mathcal{P}(s) = P(\mathcal{S} = s) = \frac{|X_s|^2}{\sum_u |X_u|^2} \tag{5}$$

where X_s is defined in (4).

In the case of OLS estimation, higher probability weight is given to subsets $\{i_1, \cdots, i_p\}$ where $\{\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_p}\}$ are more dispersed; for example, if $\boldsymbol{x}_i = (1, x_i)^T$ then for $s = \{i_1, i_2\}$, $\mathcal{P}(s) \propto (x_{i_1} - x_{i_2})^2$. For PLS estimation, the rows of the matrix L (that is, $\boldsymbol{x}_{n+1}^T, \cdots, \boldsymbol{x}_{n+r}^T$ in (2)) will contribute to the estimate $\widehat{\boldsymbol{\beta}}$ minimizing (1); in section 6, we will discuss their influence on $\widehat{\boldsymbol{\beta}}$.

The probability measure $\mathcal{P}(s)$ in (5) can also be expressed in terms of the projection matrix H defined in (3). Since

$$\sum_u |X_u|^2 = \left| \sum_{i=1}^{n+r} \boldsymbol{x}_i \boldsymbol{x}_i^T \right| = |\mathcal{X}^T \mathcal{X} + A|$$

(Mayo and Gray, 1997), it follows that

$$\begin{aligned} \mathcal{P}(s) &= \left| X_s (\mathcal{X}^T \mathcal{X} + A)^{-1} X_s^T \right| \\ &= \left| \begin{pmatrix} h_{i_1 i_1} & h_{i_1 i_2} & \cdots & h_{i_1 i_p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{i_p i_1} & h_{i_p i_2} & \cdots & h_{i_p i_p} \end{pmatrix} \right| \end{aligned}$$

where $\{h_{ij} : i, j = 1, \cdots, n+r\}$ are the elements of the projection matrix H defined in (3):

$$h_{ij} = h_{ji} = \boldsymbol{x}_i^T (\mathcal{X}^T \mathcal{X} + A)^{-1} \boldsymbol{x}_j.$$

The probability distribution $\mathcal{P}$ defined in (5) describes the weight given to each subset of p observations in defining the PLS estimate. It is also of interest to consider the total weight given to subsets of $k < p$ observations. In Proposition 1, we will show that these lower dimensional distributions depend on determinants of $k \times k$ sub-matrices of the projection matrix H defined in (3).

If $\mathcal{S}$ is a random subset of size p drawn from $\{1, \cdots, n+r\}$ with probability distribution $\mathcal{P}$, it is convenient to describe the distribution of $\mathcal{S}$ using the equivalent random vector $\boldsymbol{W} = (W_1, \cdots, W_{n+r})$ where $W_j = I(j \in \mathcal{S})$ and $W_1 + \cdots + W_{n+r} = p$. The moment

generating function $\varphi(\boldsymbol{t}) = E[\exp(\boldsymbol{t}^T\boldsymbol{W})]$ of $\boldsymbol{W}$ is given by

$$\begin{aligned}\varphi(\boldsymbol{t}) &= \left|\mathcal{X}^T\mathcal{X}+A\right|^{-1}\sum_s |X_s|^2\left\{\prod_{i_j\in s}\exp(t_{i_j})\right\}\\ &= \frac{\left|\sum_{i=1}^{n+r}\exp(t_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right|}{\left|\mathcal{X}^T\mathcal{X}+A\right|}.\end{aligned}$$

Thus for $k \le p$,

$$P(\{i_1,\cdots,i_k\}\subset\mathcal{S}) = E\left(\prod_{j=1}^k W_{i_j}\right) = \frac{\partial^k}{\partial t_{i_1}\cdots\partial t_{i_k}}\varphi(\boldsymbol{t})\bigg|_{t_1=t_2=\cdots=t_{n+r}=0}.$$

Proposition 1 below was proved in Knight (2019). For completeness, its proof can be found in the Appendix.

PROPOSITION 1. *Suppose that $\mathcal{S}$ has the distribution $\mathcal{P}$ defined in (5). Then for $k \le p$,*

$$P(\{i_1,\cdots,i_k\}\subset\mathcal{S}) = \left|\begin{pmatrix} h_{i_1i_1} & h_{i_1i_2} & \cdots & h_{i_1i_k}\\ \vdots & \vdots & \ddots & \vdots\\ h_{i_ki_1} & h_{i_ki_2} & \cdots & h_{i_ki_k}\end{pmatrix}\right|.$$

From Proposition 1, it follows that $P(W_i = 1) = h_{ii} = E(W_i)$, which suggests that an analogous measure of the influence of a subset of observations whose indices are $i_1,\cdots,i_k$ (for arbitrary $k \le n+r$) might be based on the distribution of $W_{i_1},\cdots,W_{i_k}$. This will be pursued in the next section.

Proposition 1 can be modified for the case where we minimize the penalized weighted least squares objective function

$$\sum_{i=1}^n \frac{1}{\sigma_i^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \boldsymbol{\beta}^T A\boldsymbol{\beta}$$

(for some positive $\sigma_1,\cdots,\sigma_n$) by replacing the matrix $\mathcal{X}$ in the definition of X in (2) as well as the definition of the projection matrix H in (3) by $\Sigma^{-1}\mathcal{X}$ where Σ is the diagonal matrix with diagonal elements $\sigma_1,\cdots,\sigma_n$. This modification relies on Σ being diagonal.

3 Measuring leverage for subsets

Suppose that $\mathcal{J}$ is a subset of $\{1,\cdots,n+r\}$ and given $\mathcal{S}\sim\mathcal{P}$ (as defined in (5)), define the random variable

$$N(\mathcal{J}) = \sum_{j\in\mathcal{J}} W_j = \text{card}(\mathcal{J}\cap\mathcal{S}). \tag{6}$$

Intuitively if the vectors $\{\boldsymbol{x}_j : j \in \mathcal{J}\}$ have a large influence on the PLS estimate $\widehat{\boldsymbol{\beta}}$ minimizing (1), we would expect $N(\mathcal{J})$ to be non-zero with a reasonably high probability.

It is simple to compute the mean and variance of $N(\mathcal{J})$ using Proposition 1. Applying Proposition 1 with $k = 1$ and $k = 2$, we have $E(W_i) = h_{ii}$ and $\mathrm{Var}(W_i) = h_{ii}(1 - h_{ii})$, while for $i \neq j$, $P(W_i = 1, W_j = 1) = E(W_i W_j) = h_{ii}h_{jj} - h_{ij}^2$ so that $\mathrm{Cov}(W_i, W_j) = -h_{ij}^2$. Thus

$$\begin{aligned} E[N(\mathcal{J})] &= \sum_{j\in\mathcal{J}} h_{jj} = \mathrm{tr}(H_{\mathcal{J}}) \\ \mathrm{Var}[N(\mathcal{J})] &= \sum_{j\in\mathcal{J}} h_{jj} - \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}} h_{ij}^2 = \mathrm{tr}(H_{\mathcal{J}}) - \mathrm{tr}(H_{\mathcal{J}}^2) \end{aligned}$$

where $H_{\mathcal{J}}$ is the sub-matrix of H whose row and column indices lie in $\mathcal{J}$. Note that $\mathrm{Var}[N(\mathcal{J})] = 0$ only if $H_{\mathcal{J}}$ is a projection matrix; we could use $\mathrm{Var}[N(\mathcal{J})]/E[N(\mathcal{J})] = 1 - \mathrm{tr}(H_{\mathcal{J}}^2)/\mathrm{tr}(H_{\mathcal{J}})$ as a measure of "non-projectiveness" of $H_{\mathcal{J}}$.

More generally, the probability distribution of $N(\mathcal{J})$ in (6) can be determined from its probability generating function (which is a p-th degree polynomial)

$$E\left[t^{N(\mathcal{J})}\right] = \frac{\left|t\sum_{j\in\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T + \sum_{j\notin\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T\right|}{\left|\mathcal{X}^T\mathcal{X} + A\right|}.$$

Of particular interest is $P(N(\mathcal{J}) = 0)$, which we will use to definite a measure of leverage for the set $\mathcal{J}$.

PROPOSITION 2. *If $\mathcal{J} = \{i_1, \cdots, i_k\}$ then*

$$P(N(\mathcal{J}) = 0) = \left|\begin{pmatrix} 1 - h_{i_1i_1} & -h_{i_1i_2} & \cdots & -h_{i_1i_k} \\ -h_{i_2i_1} & 1 - h_{i_2i_2} & \cdots & -h_{i_2i_k} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{i_ki_1} & -h_{i_ki_2} & \cdots & 1 - h_{i_ki_k} \end{pmatrix}\right| = |I - H_{\mathcal{J}}|$$

where $H_{\mathcal{J}}$ is the $k \times k$ sub-matrix of H whose row and column indices lie in $\mathcal{J}$.

Proof. If $\psi(t) = E\left[t^{N(\mathcal{J})}\right]$ is the probability generating function of $N(\mathcal{J})$ then $P(N(\mathcal{J}) = 0) = \psi(0)$. Define $X_{\mathcal{J}}$ to be a sub-matrix of X whose rows are $\boldsymbol{x}_{i_1}^T, \cdots, \boldsymbol{x}_{i_k}^T$ and note that

$$X_{\mathcal{J}}^T X_{\mathcal{J}} = \sum_{j\in\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T.$$

Thus

$$P\left(N(\mathcal{J}) = 0\right) = \frac{\left|\sum_{j\notin\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T\right|}{\left|\mathcal{X}^T\mathcal{X} + A\right|}$$

$$
\begin{aligned}
&= \left| I - \left(\mathcal{X}^T\mathcal{X} + A\right)^{-1} \sum_{j\in\mathcal{J}} \boldsymbol{x}_j \boldsymbol{x}_j^T \right| \\
&= \left| I - \left(\mathcal{X}^T\mathcal{X} + A\right)^{-1} X_{\mathcal{J}}^T X_{\mathcal{J}} \right| \\
&= \left| I - X_{\mathcal{J}} (\mathcal{X}^T\mathcal{X} + A)^{-1} X_{\mathcal{J}}^T \right| \\
&= \left| \begin{pmatrix} 1-h_{i_1i_1} & -h_{i_1i_2} & \cdots & -h_{i_1i_k} \\ -h_{i_2i_1} & 1-h_{i_2i_2} & \cdots & -h_{i_2i_k} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{i_ki_1} & -h_{i_ki_2} & \cdots & 1-h_{i_ki_k} \end{pmatrix} \right|,
\end{aligned}
$$

as claimed.

Proposition 2 suggests that we can use $P(N(\mathcal{J}) = 0)$ to describe the potential influence or leverage of a subset of observations $\mathcal{J}$. In particular, if $P(N(\mathcal{J}) = 0)$ is close to 1 then $\mathcal{J}$ has little influence on $\widehat{\boldsymbol{\beta}}$ minimizing (1). On the other hand, if $P(N(\mathcal{J}) = 0)$ is small (so that $P(N(\mathcal{J}) \geq 1)$ is close to 1) then the potential influence of $\mathcal{J}$ is much greater.

DEFINITION: *The leverage of a subset of observations $\mathcal{J} = \{i_1, \cdots, i_k\}$ is defined to be*

$$
lev(\mathcal{J}) = P(N(\mathcal{J}) \geq 1) = \sum_{s: s\cap\mathcal{J}\neq\emptyset} \mathcal{P}(s) = 1 - |I - H_{\mathcal{J}}|
$$

where $H_{\mathcal{J}}$ is defined in Proposition 2 and $\mathcal{P}$ is defined in (5).

Note that $\text{lev}(\mathcal{J})$ is monotone in the sense that $\mathcal{J}_1 \subset \mathcal{J}_2$ implies that $\text{lev}(\mathcal{J}_1) \leq \text{lev}(\mathcal{J}_2)$ as well as subadditive: for any $\mathcal{J}_1$ and $\mathcal{J}_2$. $\text{lev}(\mathcal{J}_1 \cup \mathcal{J}_2) \leq \text{lev}(\mathcal{J}_1) + \text{lev}(\mathcal{J}_2)$. If $\mathcal{J} = i$ (a singleton) then $\text{lev}(\mathcal{J}) = h_{ii}$, which is the standard definition of leverage for a single observation.

An alternative formula for $\text{lev}(\mathcal{J})$ is given in Proposition 3 below.

PROPOSITION 3. *If $\mathcal{J} = \{i_1, \cdots, i_k\}$ then*

$$
lev(\mathcal{J}) = 1 - \exp\left[- \int_0^1 tr\left(X_{\mathcal{J}} \left(\sum_{j\notin\mathcal{J}} \boldsymbol{x}_j \boldsymbol{x}_j^T + t \sum_{j\in\mathcal{J}} \boldsymbol{x}_j \boldsymbol{x}_j^T \right)^{-1} X_{\mathcal{J}}^T \right) dt \right].
$$

where $X_{\mathcal{J}}$ is the $k \times p$ sub-matrix of X with row indices in $\mathcal{J}$.

Proof. First of all,

$$
\begin{aligned}
|I - H_{\mathcal{J}}| &= \left| I - \left(\mathcal{X}^T\mathcal{X} + A\right)^{-1} X_{\mathcal{J}}^T X_{\mathcal{J}} \right| \\
&= \frac{\left| \mathcal{X}^T\mathcal{X} + A - X_{\mathcal{J}}^T X_{\mathcal{J}} \right|}{|\mathcal{X}^T\mathcal{X} + A|}
\end{aligned}
$$

Now define

$$g(s) = \ln\left(\left|\mathcal{X}^T\mathcal{X} + A - sX_{\mathcal{J}}^T X_{\mathcal{J}}\right|\right) - \ln\left(\left|\mathcal{X}^T\mathcal{X} + A\right|\right)$$

and note that $g(1) = \ln(|I - H_{\mathcal{J}}|)$ and $g(0) = 0$ so that

$$|I - H_{\mathcal{J}}| = \int_0^1 g'(s)\, ds.$$

Now (Golberg, 1972)

$$g'(s) = -\mathrm{tr}\left(X_{\mathcal{J}}(\mathcal{X}^T\mathcal{X} + A - sX_{\mathcal{J}}^T\mathcal{J})^{-1}X_{\mathcal{J}}^T\right)$$

and

$$\mathcal{X}^T\mathcal{X} + A - sX_{\mathcal{J}}^T X_{\mathcal{J}} = \sum_{j\notin\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T + (1-s)\sum_{j\in\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T.$$

The conclusion follows by setting $t = 1 - s$ and making a change of variables in the integral.

Proposition 3 states that $\mathrm{lev}(\mathcal{J})$ is an increasing function of

$$\int_0^1 \mathrm{tr}(D_{\mathcal{J}}(t))\, dt \tag{7}$$

where

$$D_{\mathcal{J}}(t) = X_{\mathcal{J}}\left(\sum_{j\notin\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T + t\sum_{j\in\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T\right)^{-1} X_{\mathcal{J}}^T;$$

$D_{\mathcal{J}}(1) = H_{\mathcal{J}}$. (In the case where $\mathcal{J} = i$ (a single element), $D_{\mathcal{J}}(t) = h_{ii}/(1 + (t-1)h_{ii})$.) Defining the integrand in (7) as $\omega(t) = \mathrm{tr}(D_{\mathcal{J}}(t))$, it follows that its derivatives are given by

$$\begin{aligned}
\omega'(t) &= -\mathrm{tr}(D_{\mathcal{J}}^2(t)) \le 0 \\
\omega''(t) &= 2\,\mathrm{tr}(D_{\mathcal{J}}^3(t)) \ge 0 \\
\omega^{(k)}(t) &= (-1)^k k!\,\mathrm{tr}(D_{\mathcal{J}}^{k+1}(t))
\end{aligned}$$

so that $\omega(t)$ is a non-increasing and convex function whose derivatives are monotone functions. The structure of $\omega^{(k)}(t)$ allows us to give upper and lower bounds for $\int_0^1 \omega(t)\, dt$ in terms of a small number of values of $\omega(t)$. For example, since $\omega^{(4)}(t) \ge 0$, it follows from Bullen (1978) that

$$\frac{3}{8}\omega(1/6) + \frac{1}{4}\omega(1/2) + \frac{3}{8}\omega(5/6) \le \int_0^1 \omega(t)\, dt \le \frac{1}{8}\omega(0) + \frac{3}{8}\omega(1/3) + \frac{3}{8}\omega(2/3) + \frac{1}{8}\omega(1).$$

Related upper and lower bounds can be found in Bessenyei and Páles (2002).

The sub-matrix $H_{\mathcal{J}}$ has been used to define a number of measures for quantifying the leverage of a subset of observations in the case of OLS estimation. Barrett and Gray (1997) define the leverage of a subset of observations $\mathcal{J}$ to be the Frobenius norm of $H_{\mathcal{J}}$ while Cook

and Weisberg (1982) define the leverage of $\mathcal{J}$ as the maximum eigenvalue of $H_{\mathcal{J}}$. Note that if $\mu_1, \cdots, \mu_k$ are the eigenvalues of $H_{\mathcal{J}}$ then

$$\text{lev}(\mathcal{J}) = 1 - \prod_{j=1}^{k}(1-\mu_j) \geq \max_{1\leq j\leq k} \mu_j$$

with equality only if the maximum eigenvalue is 1 or if $k-1$ of the eigenvalues are 0. An example of the latter situation occurs when $\mathcal{J} = \{i_1, \cdots, i_k\}$ with $\boldsymbol{x}_{i_1} = \cdots = \boldsymbol{x}_{i_k}$, in which case $H_{\mathcal{J}}$ has one eigenvalue equal to $kh_{i_1 i_1} = h_{i_1 i_1} + \cdots + h_{i_k i_k}$ and $k-1$ eigenvalues equal to 0.

The Cook and Weisberg (1982) definition of leverage has a simple geometric interpretation in the case where $\mathcal{J} = \{i_1, \cdots, i_k\} \subset \{1, \cdots, n\}$. Define $\mathcal{E}_{\mathcal{J}}$ to be the space spanned by the coordinate vectors whose indices lie in $\mathcal{J}$. For $\boldsymbol{v} \in \mathcal{E}_{\mathcal{J}}$, we can compare $H\boldsymbol{y}$ and $H(\boldsymbol{y}+\boldsymbol{v})$: Clerc Bérod and Morgenthaler (1997) show that

$$\frac{\|H(\boldsymbol{y}+\boldsymbol{v}) - H\boldsymbol{y}\|^2}{\|\boldsymbol{v}\|^2} = \frac{\boldsymbol{v}^T H \boldsymbol{v}}{\|\boldsymbol{v}\|^2} = \cos^2(\theta(\boldsymbol{v}))$$

where $\theta(\boldsymbol{v})$ is the angle between $\boldsymbol{v} \in \mathcal{E}$ and $H\boldsymbol{v}$. The angle $\theta(\boldsymbol{v})$ is minimized (and $\cos^2(\theta(\boldsymbol{v}))$ maximized) when $\cos^2(\theta(\boldsymbol{v}))$ is equal to the maximum eigenvalue of $H_{\mathcal{J}}$.

Similarly, our definition of leverage can be interpreted geometrically in the case where $\mathcal{J} = \{i_1, \cdots, i_k\} \subset \{1, \cdots, n\}$. Our approach differs in the sense that we compare "residuals" $(I-H)\boldsymbol{y}$ and $(I-H)(\boldsymbol{y}+\boldsymbol{v})$ for $\boldsymbol{v} \in \mathcal{E}_{\mathcal{J}}$; if $\boldsymbol{v} \in \mathcal{E}_{\mathcal{J}}$ has a large effect on the fitted values (for indices in $\mathcal{J}$), its effect on the residuals will be small. Suppose that V is a subset of $\mathcal{E}_{\mathcal{J}}$ of the form

$$V = V(U) = \{\boldsymbol{v} : (v_{i_1}, \cdots, v_{i_k}) \in U \text{ and } v_j = 0 \text{ for } j \notin \mathcal{J}\}$$

where volume$(U) > 0$. For $\boldsymbol{v} \in V$, define

$$\boldsymbol{a}(\boldsymbol{v}) = (I-H)(\boldsymbol{y}+\boldsymbol{v}) - (I-H)\boldsymbol{y} = (I-H)\boldsymbol{v}$$

and

$$D = \{(a_{i_1}(\boldsymbol{v}), \cdots, a_{i_k}(\boldsymbol{v})) : \boldsymbol{v} \in V\}.$$

Then lev$(\mathcal{J})$ is related to the volume reduction from U to D:

$$\frac{\text{volume}(D)}{\text{volume}(U)} = |I - H_{\mathcal{J}}| = 1 - \text{lev}(\mathcal{J}).$$

This geometric interpretation also applies to the case where $\widehat{\boldsymbol{y}}_0 = S\boldsymbol{y}_0$ where S is a smoothing matrix; using the argument given above, we can define for a subset $\mathcal{J} \subset \{1, \cdots, n\}$,

$$\text{lev}(\mathcal{J}) = 1 - |I - S_{\mathcal{J}}|$$

where $S_{\mathcal{J}}$ is the sub-matrix of S whose row and columns indices are in $\mathcal{J}$ and $|I - S_{\mathcal{J}}|$ is the absolute value of the determinant. In section 5, we will consider the special case where S is symmetric with eigenvalues in $[0, 1]$.

Depending on the size of the set $\mathcal{J}$, $\text{lev}(\mathcal{J})$ can be approximated in a number of ways. If $H_{\mathcal{J}}$ has eigenvalues all less than 1 then we can expand $\ln(|I - H_{\mathcal{J}}|)$ as follows:

$$\ln(|I - H_{\mathcal{J}}|) = -\sum_{k=1}^{\infty} \frac{\text{tr}(H_{\mathcal{J}}^k)}{k}.$$

Thus, for example, when $h_{i_1 i_1}, \cdots, h_{i_k i_k}$ are uniformly small and $k \ll n$ then

$$P(N(\mathcal{J}) = 0) \approx \exp\left(-\sum_{j=1}^{k} h_{i_j i_j} - \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2\right).$$

and so

$$\text{lev}(\mathcal{J}) \approx 1 - \exp\left(-\sum_{j=1}^{k} h_{i_j i_j} - \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2\right).$$

There is a simple connection between this approximation and the result of Proposition 3. For t close to 1, if $\mathcal{J} = \{i_1, \cdots, i_k\}$ then

$$\begin{aligned}
&\text{tr}\left(X_{\mathcal{J}}\left(\sum_{j\notin\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T + t\sum_{j\in\mathcal{J}} \boldsymbol{x}_j\boldsymbol{x}_j^T\right)^{-1} X_{\mathcal{J}}^T\right) \\
&\quad\approx\quad \text{tr}\left(X_{\mathcal{J}}(X^TX)^{-1}X_{\mathcal{J}}^T\right) - (t-1)\text{tr}\left(\left\{X_{\mathcal{J}}(X^TX)^{-1}X_{\mathcal{J}}^T\right\}^2\right) \\
&\quad=\quad \sum_{j=1}^{k} h_{i_j i_j} - (t-1)\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2
\end{aligned}$$

and

$$\int_0^1 \left\{\sum_{j=1}^{k} h_{i_j i_j} - (t-1)\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2\right\} dt = \sum_{j=1}^{k} h_{i_j i_j} + \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2.$$

On the other hand, if the set of indices $\mathcal{J}$ is large then $\text{lev}(\mathcal{J})$ can be estimated using Monte Carlo sampling. Using the fact that $\text{tr}(B) = E(\boldsymbol{Z}^T B\boldsymbol{Z})$ for any random vector $\boldsymbol{Z}$ with $E(\boldsymbol{Z}\boldsymbol{Z}^T) = I$, it follows that

$$\ln(|I - H_{\mathcal{J}}|) = -\sum_{k=1}^{\infty} \frac{\text{tr}(H_{\mathcal{J}}^k)}{k} = -\sum_{k=1}^{\infty} \frac{E(\boldsymbol{Z}^T H_{\mathcal{J}}^k \boldsymbol{Z})}{k}$$

if all the eigenvalues of $H_{\mathcal{J}}$ are less than 1. (Likewise we can write the maximum eigenvalue of $H_{\mathcal{J}}$ as

$$\ln\left(\max_{1\le j\le k} \mu_j\right) = \lim_{k\to\infty} \frac{\ln[E(\boldsymbol{Z}^T H_{\mathcal{J}}^k \boldsymbol{Z})]}{k}$$

where $\mu_1, \cdots, \mu_k$ are the eigenvalues of $H_{\mathcal{J}}$ and $E(\boldsymbol{Z}\boldsymbol{Z}^T) = I$.) We can then use the Hutchinson-Skilling method (Hutchinson, 1990; Skilling, 1989) to estimate $\text{tr}(H_{\mathcal{J}}^k)$:

$$\widehat{\text{tr}}(H_{\mathcal{J}}^k) = \frac{1}{m}\sum_{t=1}^{m} \boldsymbol{Z}_t^T P_{\mathcal{J}}(HP_{\mathcal{J}})^k \boldsymbol{Z}_t$$

for some m where $P_{\mathcal{J}}$ is a projection onto $\mathcal{E}_{\mathcal{J}}$, the space spanned by the coordinate vectors whose indices are in $\mathcal{J}$ (so that $H^k_{\mathcal{J}}$ is the sub-matrix of $P_{\mathcal{J}}(HP_{\mathcal{J}})^k$ whose row and column indices are in $\mathcal{J}$ with the remaining elements of $P_{\mathcal{J}}(HP_{\mathcal{J}})^k$ equal to 0) and $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_m$ are independent random vectors with $E[\boldsymbol{Z}_t \boldsymbol{Z}_t^T] = I$. In practice, we typically take $\{\boldsymbol{Z}_t\}$ to have components that are independent Rademacher random variables taking values ± 1 each with probability 1/2.

The result of Proposition 3 can also be used to approximate $\text{lev}(\mathcal{J})$ using some sort of numerical quadrature to approximate the integral. The downside of this approach is the need to compute QR decompositions at multiple values of $t \in (0, 1)$ in order to approximate $\int_0^1 \text{tr}(D_{\mathcal{J}}(t))\, dt$; $\text{tr}(tD_{\mathcal{J}}(t))$ is the sum of diagonal elements of a projection matrix, which can obtained from sum of the squared L_2 norms of the Q matrix in the QR decomposition. Drineas *et al.* (2011, 2012) give algorithms for computationally efficient approximations of the diagonal elements of a projection matrix.

4 Example: The Hodrick-Prescott filter

Consider the Hodrick-Prescott (H-P) filter (Hodrick and Prescott, 1997) for smoothing a time series $\{y_t\}$; for some $\lambda > 0$, we define $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$ to minimize

$$\sum_{t=1}^{n} (y_t - \theta_t)^2 + \lambda \sum_{t=2}^{n-1} (\theta_{t+1} - 2\theta_t + \theta_{t-1})^2$$

The H-P filter is a low-pass filter and is closely related to the method of "graduation"in actuarial science proposed by Whittaker (1922).

In economics, the H-P filter is often used to decompose the time series $\{y_t\}$ into trend (represented by $\{\theta_t\}$) and cyclical components. By varying λ, we can vary the degree of low-pass filtering: As λ increases, the penalty term (which penalizes the squared second differences) becomes more dominant and so the output of the H-P filter becomes smoother. A fast algorithm for computing $\{\widehat{\theta}_t\}$ is given in Cornea-Madeira (2017). Hamilton (2018) gives a compelling critique of this method, noting its drawbacks for many economic time series.

Figures 1 and 2 show the leverages for each row of the matrices $\mathcal{X} = I$ and L (whose $n-2$ rows contain the vector $(\sqrt{\lambda} \quad -2\sqrt{\lambda} \quad \sqrt{\lambda})$ surrounded by $n-3$ zeroes) for $\lambda = 1600$ (which is often used as a default value for quarterly economic data) and $\lambda = 100$ with $n = 200$ for both values of λ. What is notable in both cases is the influence of the rows of $\mathcal{X}$ and L at the two ends of the data. For example, note that the leverages of the first and last rows of L are very close to 1 and the leverages of the first and last rows of $\mathcal{X}$ are much larger than those of the middle rows. We can also compute the leverage of subsets: For $\lambda = 1600$ ($\lambda = 100$) and $n = 200$, the leverage of the first 10 rows of $\mathcal{X}$ is 0.893 (0.989) compared to 0.525 (0.843) for rows 91-100. This suggests that, at the ends of the data, the values of $\{\widehat{\theta}_t\}$ are estimated by

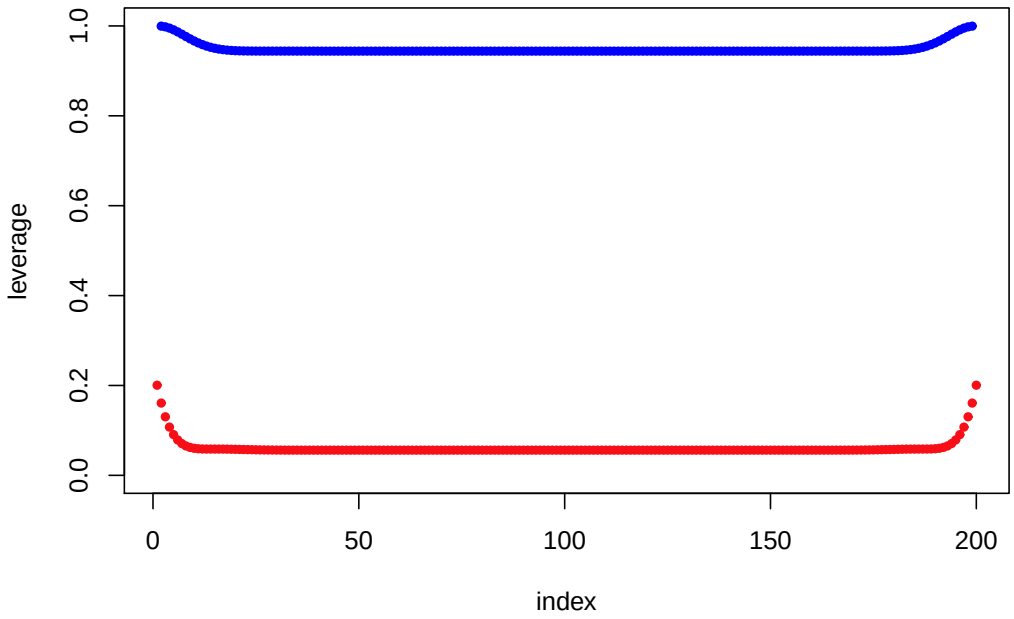

Figure 1: Leverage for each row of $\mathcal{X}$ (red) and each row of L (blue) for the Hodrick-Prescott filter with $n = 200$ and $\lambda = 1600$.

a relatively smaller number of $\{y_t\}$ and so the nature of the H-P filter is quite different for these data. Indeed Hamilton (2018) observes: "Filtered values at the end of the sample are very different from those in the middle and are also characterized by spurious dynamics."

If we have the infinite past and future of the process $\{y_t\}$ then we can derive the form of $\{\widehat{\theta}_t\}$ using frequency domain methods:

$$\widehat{\theta}_t = \sum_{u=-\infty}^{\infty} c_u y_{t-u}$$

where $\{c_u\}$ satisfy

$$c_u = \int_0^1 \frac{\cos(2\pi us)}{1 + 16\lambda \sin^4(\pi s)}\, ds.$$

From this, it follows that the leverage of a row of the infinite dimensional $\mathcal{X}$ is

$$\chi(\lambda) = \int_0^1 \frac{1}{1 + 16\lambda \sin^4(\pi s)}\, ds$$

while the leverage of a row of the corresponding L is

$$1 - \chi(\lambda) = \int_0^1 \frac{16\lambda \sin^4(\pi s)}{1 + 16\lambda \sin^4(\pi s)}\, ds.$$

In the examples given above (for $n = 200$), these values are close to the leverage values for the middle observations.

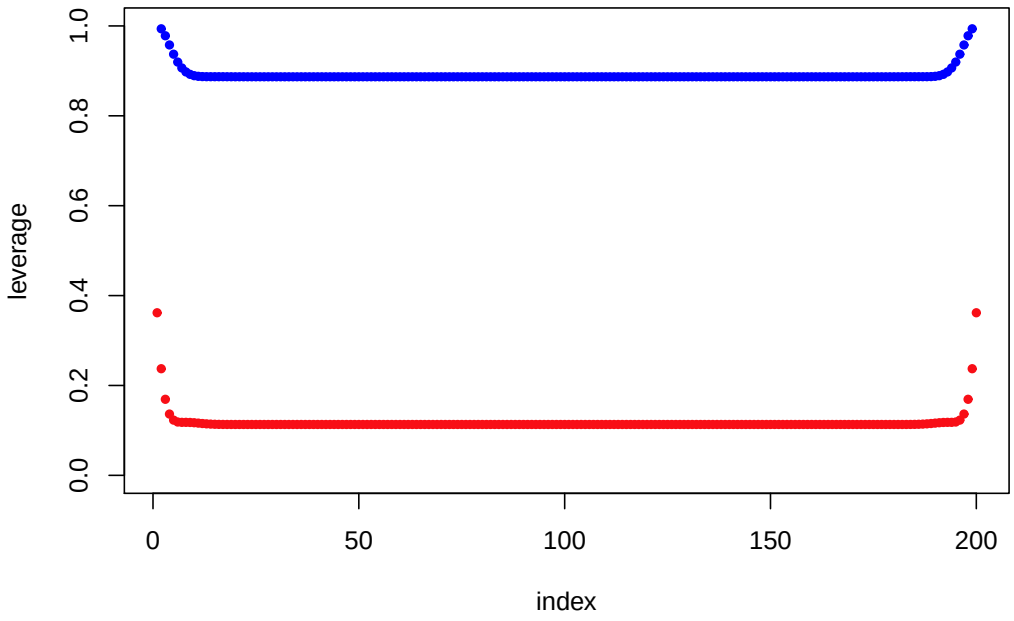

Figure 2: Leverage for each row of $\mathcal{X}$ (red) and each row of L (blue) for the Hodrick-Prescott filter with $n = 200$ and $\lambda = 100$.

An obvious resolution to the issue above is to downweight the observations near the ends of the data in order to make their influence similar to the remaining observations. For example, we could define $\{\widehat{\theta}_t\}$ to minimize

$$\sum_{t=1}^{n} w_t(y_t - \theta_t)^2 + \lambda \sum_{t=2}^{n-1} v_t(\theta_{t+1} - 2\theta_t + \theta_{t-1})^2$$

for some non-negative weights $\{w_t\}$ and $\{v_t\}$ with $w_1 + \cdots + w_n + v_2 + \cdots + v_{n-1} = 2n - 2$. One possible approach is to define the weights so that the leverages of the corresponding $\mathcal{X}$ and L are proportional to $\chi(\lambda)$ and $1 - \chi(\lambda)$, respectively.

As an illustration, suppose that $\lambda = 1600$. Then $\chi(\lambda) = 0.05608$. Figure 3 shows the weights $\{w_t\}$ and $\{v_t\}$ for $n = 200$ (and $\lambda = 1600$). The weights for the end observations are extremely small relative to those for the middle observations. While this may seem somewhat counter-intuitive, it should be remembered that this weighted H-P filter, like its unweighted counterpart, is essentially a local smoothing method; thus even though the weights for the end observations are small (relative to the middle observation weights), they are not excessively small relative to each other. Figure 4 shows trend estimates for the H-P filter and its "equi-leverage" version for a simulated time series of length $n = 200$ using $\lambda = 1600$; in the middle of the data, the two estimates of trend are qualitative similar but are quite different at the two ends, with the H-P filter producing somewhat smoother estimates

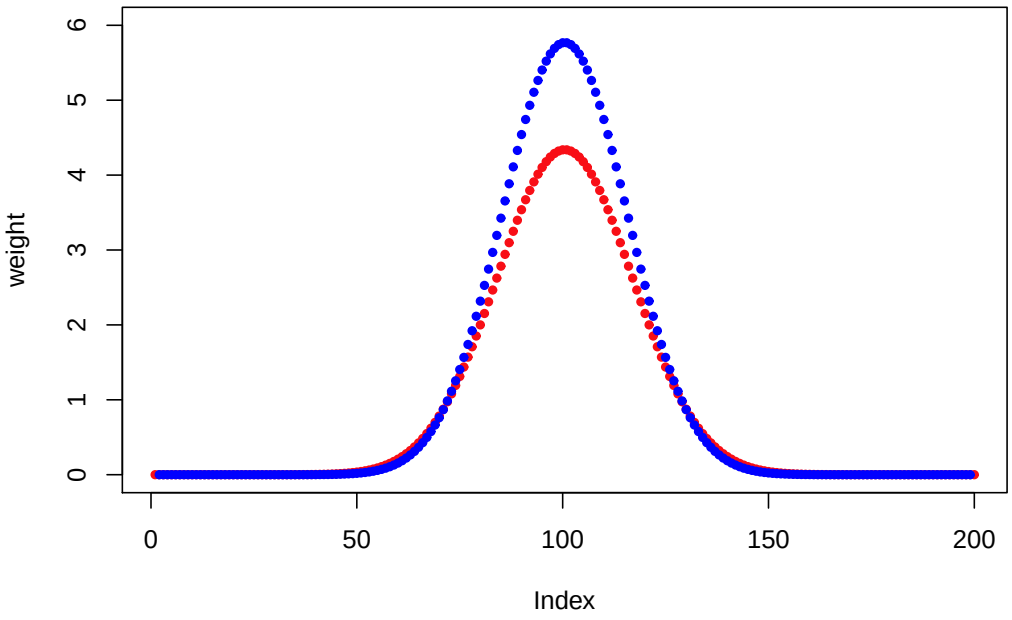

Figure 3: Weights $\{w_t : t = 1, \cdots, 200\}$ (red) and $\{v_t : t = 2, \cdots, 199\}$ (blue) for $n = 200$ and $\lambda = 1600$.

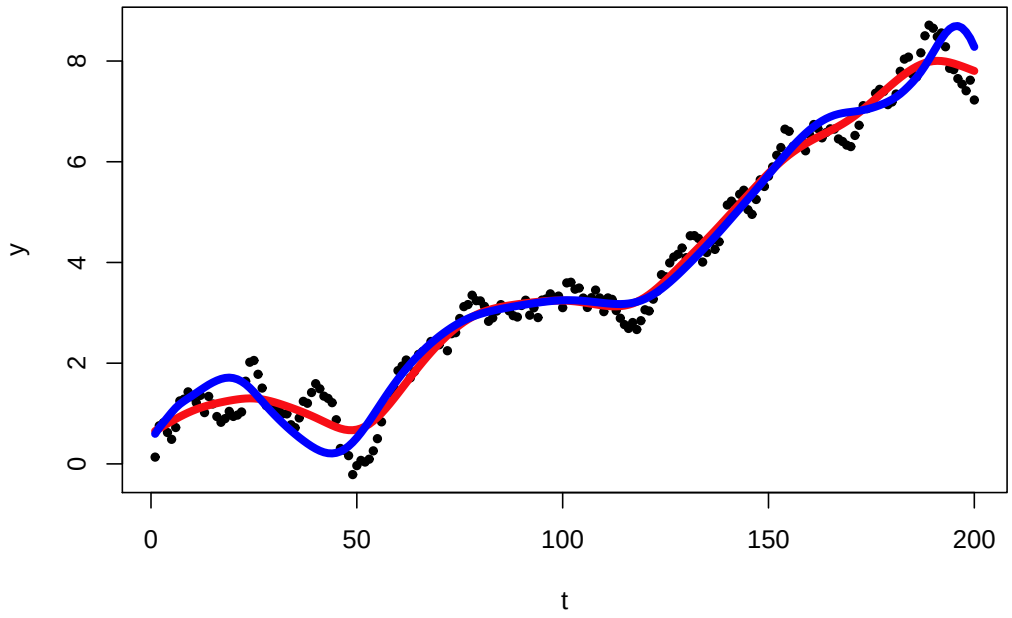

Figure 4: Simulated data with H-P filter (red) and equi-leverage version (blue).

than its equi-leverage cousin.

5 Example: Smoothing matrices and backfitting

Suppose that S is an $n \times n$ symmetric matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$ lying in the interval $[0, 1]$ and define

$$\widehat{\boldsymbol{y}}_0 = S\boldsymbol{y}_0 \tag{8}$$

where (as before) $\boldsymbol{y}_0 = (y_1, \cdots, y_n)^T$. The matrix S is a special case of a smoothing matrix. (More generally, a smoothing matrix may be asymmetric with eigenvalues lying in $(-1, 1]$.) In practice, the smoothing matrix S is typically not explicitly evaluated (and may be very difficult to evaluate) although its properties (for example, trace and eigenstructure) are usually easier to compute, either exactly or approximately. For example, many ensemble methods such as bagging (Breiman, 1996) and stacking (Wolpert, 1992) combine OLS estimates from several models so that $S = a_1 H_1 + \cdots + a_r H_r$ for positive $a_1, \cdots, a_r$ and projection matrices $H_1, \cdots, H_r$; in fact, Choi and Wu (1990) show that S can also be expressed as a convex combination of r projection matrices with $r \leq \lfloor \log_2(n) \rfloor + 2$.

Suppose that $\lambda_1, \cdots, \lambda_m$ ($m \leq n$) are the non-zero eigenvalues of S with $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m$ the corresponding orthonormal eigenvectors and define Γ to be the matrix whose columns are $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m$; the vector $\widehat{\boldsymbol{y}}_0$ in (8) can be written as

$$\widehat{\boldsymbol{y}}_0 = \sum_{j=1}^{m} \widehat{\beta}_j \boldsymbol{v}_j = \Gamma \widehat{\boldsymbol{\beta}}$$

where $\{\widehat{\beta}_j\}$ are PLS estimates minimizing

$$\|\boldsymbol{y}_0 - \Gamma\boldsymbol{\beta}\|^2 + \sum_{j=1}^{m} \frac{(1-\lambda_j)}{\lambda_j} \beta_j^2.$$

To see this, define Λ to be the $m \times m$ diagonal matrix whose elements are $\lambda_1, \cdots, \lambda_m$ and note that $S = \Gamma\Lambda\Gamma^T$; the matrix X in (2) is given by

$$X = \begin{pmatrix} \Gamma \\ \Lambda^{-1/2}(I - \Lambda)^{1/2} \end{pmatrix}$$

with $X^T X = \Lambda^{-1}$. Then the matrix H in (3) becomes

$$H = \begin{pmatrix} \Gamma\Lambda\Gamma^T & \Gamma\Lambda^{1/2}(I-\Lambda)^{1/2} \\ \Lambda^{1/2}(I-\Lambda)^{1/2}\Gamma^T & I - \Lambda \end{pmatrix} = \begin{pmatrix} S & \Gamma\Lambda^{1/2}(I-\Lambda)^{1/2} \\ \Lambda^{1/2}(I-\Lambda)^{1/2}\Gamma^T & I - \Lambda \end{pmatrix}$$

and so

$$H\boldsymbol{y} = \begin{pmatrix} S\boldsymbol{y}_0 \\ \Lambda^{1/2}(I-\Lambda)^{1/2}\Gamma^T\boldsymbol{y}_0 \end{pmatrix}.$$

Note that the standard definition of equivalent degrees of freedom of the linear smoother given by S is $\text{tr}(S) = E[N(\mathcal{J})]$ where $\mathcal{J} = \{1, \cdots, n\}$.

Backfitting is a commonly used method for estimating (non-parametrically) the functions $f_1, \cdots, f_p$ in the additive model

$$Y_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i \quad (i = 1, \cdots, n).$$

Given data $\{(\boldsymbol{x}_i, y_i)\}$, we can estimate $f_1, \cdots, f_p$ using the backfitting algorithm, which iteratively smooths the data to produce estimates of $f_1, \cdots, f_p$. Defining

$$\mathbf{f}_j = \begin{pmatrix} f_j(x_{1j}) \\ \vdots \\ f_j(x_{nj}) \end{pmatrix} \quad (j = 1, \cdots, p)$$

we estimate $\mathbf{f}_1, \cdots, \mathbf{f}_p$ iteratively as follows:

$$\widehat{\mathbf{f}}_j \leftarrow S_j(\boldsymbol{y}_0 - \sum_{k \neq j} \widehat{\mathbf{f}}_k)$$

where $S_1, \cdots, S_p$ are $n \times n$ smoothing matrices. If the backfitting algorithm converges then $\{\widehat{\mathbf{f}}_j\}$ (at convergence) satisfy

$$\begin{pmatrix} I & S_1 & S_1 & \cdots & S_1 & S_1 \\ S_2 & I & S_2 & \cdots & S_2 & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ S_p & S_p & S_p & \cdots & S_p & I \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{f}}_1 \\ \widehat{\mathbf{f}}_2 \\ \vdots \\ \widehat{\mathbf{f}}_p \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_p \end{pmatrix} \boldsymbol{y}_0.$$

Thus we can write $\widehat{\boldsymbol{y}}_0 = S\boldsymbol{y}_0$ where

$$S = (I \; I \cdots I) \begin{pmatrix} I & S_1 & S_1 & \cdots & S_1 & S_1 \\ S_2 & I & S_2 & \cdots & S_2 & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ S_p & S_p & S_p & \cdots & S_p & I \end{pmatrix}^{-1} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_p \end{pmatrix}$$

where the matrix inverse above exists if $\|S_j S_k\| < 1$ for some j and k (and for some matrix norm). It can also be shown that S is symmetric if $S_1, \cdots, S_p$ are symmetric. In practice, explicit evaluation of S is difficult although we can use techniques such as the Hutchinson-Skilling method to compute leverage for subsets of observations.

As an illustration, suppose that we want to estimate the functions g_1 and g_2 in the model

$$Y_i = f_1(x_{1i}) + f_2(x_{2i}) + \varepsilon_i \quad \text{for } i = 1, \cdots, n = 1000$$

where f_1 and f_2 are estimated using smoothing splines with 5 (equivalent) degrees of freedom. The points $\{x_{1i}\}$ and $\{x_{2i}\}$ are both uniformly distributed over the set $\{k/1001 : k =$

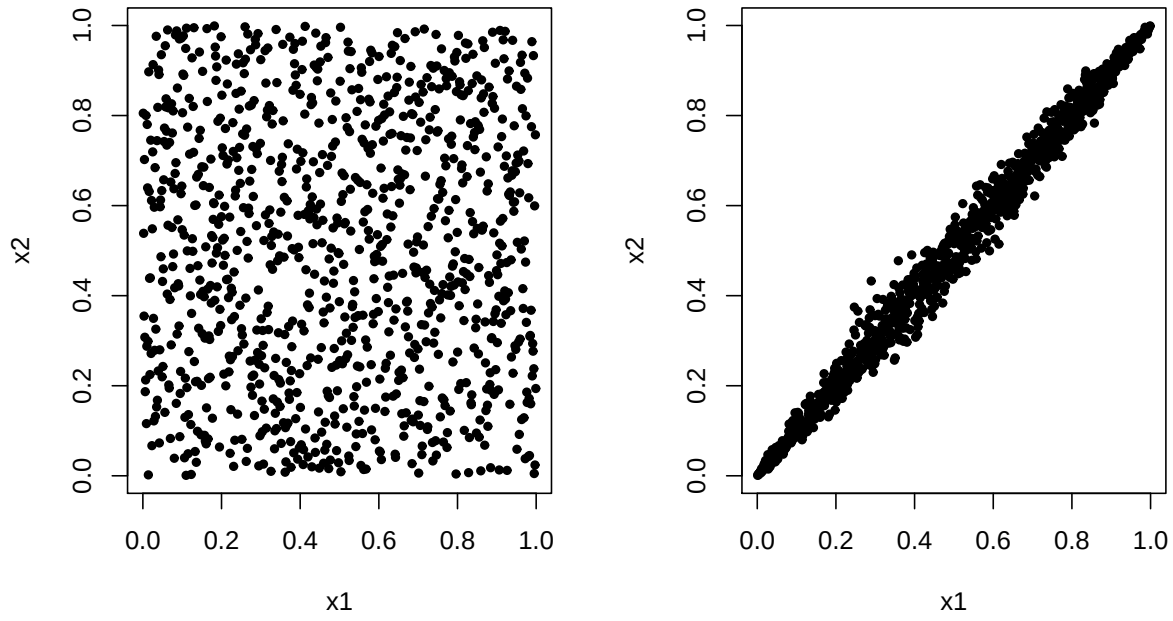

Figure 5: Configuration of points $\{(x_{i1}, x_{i2})\}$ for the two designs; the correlations are 0.01 and 0.99, respectively.

$1, \cdots, 1000\}$ and we consider two designs for $\{(x_{1i}, x_{2i})\}$ (shown in Figure 5) – one with low correlation and one with high correlation. We define S_1 and S_2 so that S_1 corresponds to a smoothing spline with 5 (equivalent) degrees of freedom and S_2 is "centered" so that $S_2\mathbf{1} = \mathbf{0}$ where $\mathbf{1}$ is a vector of 1s; thus $\mathrm{tr}(S_1) = 5$ and $\mathrm{tr}(S_2) = 4$. The matrix S is given by

$$S = S_1 + (I - S_1)(I - S_2S_1)^{-1}S_2(I - S_1) = S_2 + (I - S_2)(I - S_1S_2)^{-1}S_1(I - S_2).$$

For the low correlation design in Figure 5 (where the correlation is 0.01), we have $S_1S_2 \approx 0$ so that $\mathrm{tr}(S) \approx \mathrm{tr}(S_1) + \mathrm{tr}(S_2)$ and in fact, $\mathrm{tr}(S) = 8.99$. For the high correlation design (where the correlation is 0.99), we have $S_1S_2 \approx S_2^2 \approx S_2S_1$ and

$$\begin{aligned}
\mathrm{tr}(S) &= \mathrm{tr}(S_1) + \sum_{j=0}^{\infty} \mathrm{tr}\left((I - S_1)(S_2S_1)^j S_2(I - S_1)\right) \\
&= \mathrm{tr}(S_1) + \sum_{j=0}^{\infty} \mathrm{tr}\left(S_2(S_2S_1)^j(I - S_1)^2\right) \\
&\approx \mathrm{tr}(S_1) + \mathrm{tr}(S_2) - 2\,\mathrm{tr}(S_2^2) + \mathrm{tr}(S_2^3) \\
&\quad + \sum_{j=1}^{\infty} \mathrm{tr}\left(S_2^{2j+1} - 2S_2^{2j+2} + S_2^{2j+3}\right)
\end{aligned}$$

For the high correlation design in Figure 5, $\mathrm{tr}(S) = 6.18$.

Table 1 gives estimated values of $\text{lev}(\mathcal{J})$ for $\mathcal{J} = \{i : (k-1)/10 < x_{1i} < k/10\}$ where $k = 1, \cdots, 5$; these estimates are computed using the Hutchinson-Skilling method as described in section 3. The leverages for the low correlation design are higher than those for the high correlation design as one would expect given the greater equivalent degrees of freedom for the low correlation design.

	Range of x_1				
Design	$(0.0, 0.1)$	$(0.1, 0.2)$	$(0.2, 0.3)$	$(0.3, 0.4)$	$(0.4, 0.5)$
low correlation	0.80	0.60	0.58	0.57	0.58
high correlation	0.77	0.45	0.44	0.43	0.44

Table 1: Estimated leverage for subsets of observations whose x_1 value lies in the given range. The standard errors of these estimates are between 0.004 and 0.007.

To put the numbers in Table 1 into some context, suppose that we assume that f_1 and f_2 are quartic polynomials and estimate the 9 coefficients using OLS for the low correlation design. In this case, we can evaluate the leverages for the $\mathcal{J}$ in Table 1 exactly; these are given in Table 2. Note that the OLS leverages for the more extreme values of x_1 are substantially greater than those for backfitting while the leverages for the more central values of x_1 are similar.

	Range of x_1				
Method	$(0.0, 0.1)$	$(0.1, 0.2)$	$(0.2, 0.3)$	$(0.3, 0.4)$	$(0.4, 0.5)$
Backfitting	0.80	0.60	0.58	0.57	0.58
Quartic OLS	0.95	0.66	0.59	0.55	0.58

Table 2: Leverages of backfitting and quartic OLS for the low correlation design.

6 Influence of the penalty term

The leverage measure defined in section 3 can be used to investigate the influence of the penalty term $\boldsymbol{\beta}^T A \boldsymbol{\beta}$ on the PLS estimate $\widehat{\boldsymbol{\beta}}$. This is of particular interest when p is much smaller than n.

Define $\mathcal{J}_{data} = \{1, \cdots, n\}$ and $\mathcal{J}_{pen} = \{n+1, \cdots, n+r\}$; the forms of both $\text{lev}(\mathcal{J}_{data})$ and $\text{lev}(\mathcal{J}_{pen})$ are similar. First of all,

$$\begin{aligned} \text{lev}(\mathcal{J}_{pen}) &= 1 - \left| I - L(\mathcal{X}^T\mathcal{X} + A)^{-1} L^T \right| \\ &= 1 - \left| I - (\mathcal{X}^T\mathcal{X} + A)^{-1} L^T L \right| \end{aligned}$$

$$
\begin{aligned}
&= 1 - \left| I - (\mathcal{X}^T\mathcal{X} + A)^{-1}(A + \mathcal{X}^T\mathcal{X} - \mathcal{X}^T\mathcal{X}) \right| \\
&= 1 - \left| (\mathcal{X}^T\mathcal{X} + A)^{-1}\mathcal{X}^T\mathcal{X} \right| \\
&= 1 - \frac{|\mathcal{X}^T\mathcal{X}|}{|\mathcal{X}^T\mathcal{X} + A|}
\end{aligned}
$$

In general, $\text{lev}(\mathcal{J}_{pen})$ depends only on A and not L. Note that $\text{lev}(\mathcal{J}_{pen}) = 0$ if, and only if, $A = 0$. On the other hand, if $p \approx n$ then $\text{lev}(\mathcal{J}_{pen})$ is often very close to 1. For example, for the H-P filter discussed in section 4, $\text{lev}(\mathcal{J}_{pen})$ is effectively 1 for both $\lambda = 100$ and $\lambda = 1600$ when $n = 200$. In the case of a symmetric non-negative definite smoothing matrix S (discussed in section 5),

$$\text{lev}(\mathcal{J}_{pen}) = 1 - \prod_{j=1}^{m} \lambda_j$$

where $\lambda_1, \cdots, \lambda_m$ are the non-zero eigenvalues of S.

For $\text{lev}(\mathcal{J}_{data})$, we get

$$\text{lev}(\mathcal{J}_{data}) = 1 - \frac{|A|}{|\mathcal{X}^T\mathcal{X} + A|}$$

where the derivation follows similarly to that for $\text{lev}(\mathcal{J}_{pen})$. From this, it follows that $\text{lev}(\mathcal{J}_{data}) = 1$ unless A has rank p and hence has non-zero determinant. The latter case occurs when the rank of $\mathcal{X}$ is less than p (so that OLS estimates are not unique) and the penalty is used to impose constraints on the parameters. An example of this is spline smoothing where for cubic splines, there are $n + 2$ parameters and the rank of $\mathcal{X}$ is n; see, for example, Hastie and Tibshirani (1990) for more details.

From a data analytic point of view, $\text{lev}(\mathcal{J}_{pen})$ is potentially interesting as it reflects the influence of the penalty term on the estimates. As most PLS estimates depend on a tuning parameter (or parameters), it is possible to adjust these to control or limit the influence of the penalty term. In the next section, we will illustrate this in the context of ridge regression.

7 Example: Ridge regression

Ridge regression (Hoerl and Kennard, 1970) is often used in the case where the regression design has a high degree of (near-) collinearity, that is, when one or more of the eigenvalues of $\mathcal{X}^T\mathcal{X}$ is small or even 0. The idea behind ridge regresson is to shrink OLS estimates towards 0, which increases their absolute bias while simultaneously reducing their variance. The shrinkage is controlled by a tuning parameter; the goal is to find a value of the tuning parameter that minimizes or nearly minimizes the mean square error.

In ridge regression, the standard practice is to normalize the predictors to have mean 0 and equal variances. This allows us to estimate the intercept parameter by $\bar{y}$, the average of $y_1, \cdots, y_n$; thus by subtracting $\bar{y}$ from $y_1, \cdots, y_n$, we can remove the intercept parameter from the model. In addition, we will scale the predictors $\{\boldsymbol{x}_i : i = 1, \cdots, n\}$ are so that the

diagonal elements of the matrix $\mathcal{X}^T\mathcal{X}$ are all 1; thus $\mathcal{X}^T\mathcal{X}$ is the correlation matrix of the p predictors. $\mathcal{X}^T\mathcal{X}$ need not be invertible, for example, if $p > n$.

The ridge estimate $\widehat{\boldsymbol{\beta}}_\lambda$ is defined as the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

where $\lambda \geq 0$ is a tuning parameter that controls the shrinkage of the OLS estimates. From above, it follows that the leverage of the penalty (setting $A = \lambda I$ with $L = \lambda^{1/2}I$) is given by

$$\text{lev}(\mathcal{J}_{pen}) = 1 - \frac{|\mathcal{X}^T\mathcal{X}|}{|\mathcal{X}^T\mathcal{X} + \lambda I|} = 1 - \prod_{j=1}^{p}\left(\frac{\mu_j}{\mu_j + \lambda}\right)$$

where $\mu_1, \cdots, \mu_p$ are the eigenvalues of $\mathcal{X}^T\mathcal{X}$. From Jensen's inequality, we have

$$\text{lev}(\mathcal{J}_{pen}) \geq 1 - (1 + \lambda)^{-p}$$

with equality if, and only if, $\mu_1 = \cdots = \mu_p = 1$ (in which case, the predictors are uncorrelated). As illustrated in Dobriban and Wager (2018), the eigenvalues of $\mathcal{X}^T\mathcal{X}$ play an important role in determining the optimal prediction mean square error.

As one might expect (and perhaps hope), $\text{lev}(\mathcal{J}_{pen})$ increases as $\mu_1, \cdots, \mu_p$ become more dispersed – the influence of the penalty increases as the design becomes more collinear. When $p > n$ then at least one eigenvalue of $\mathcal{X}^T\mathcal{X}$ equals 0 and so $\text{lev}(\mathcal{J}_{pen}) = 1$; in other words, every elemental estimate with a non-zero weight in $\widehat{\boldsymbol{\beta}}_\lambda$ uses at least one of $(\boldsymbol{x}_{n+1}, 0), \cdots, (\boldsymbol{x}_{n+p}, 0)$. (An elemental estimate using $(\boldsymbol{x}_{n+i_1}, 0), \cdots, (\boldsymbol{x}_{n+i_k}, 0)$ will produce zeroes in the $i_1, \cdots, i_k$ components of the elemental estimate.)

Now consider a more concrete (albeit artificial) example. Suppose $\mathcal{X}^T\mathcal{X} = R_\gamma$ where all the off-diagonal elements are equal to γ for some $\gamma \in [0, 1)$. The eigenvalues of R_γ are $1 + (p-1)\gamma$ and $1 - \gamma$ (with multiplicity $p - 1$) From this, we get

$$\text{lev}(\mathcal{J}_{pen}) = 1 - \left(\frac{1 + (p-1)\gamma}{1 + (p-1)\gamma + \lambda}\right)\left(\frac{1 - \gamma}{1 - \gamma + \lambda}\right)^{p-1}$$

and

$$\frac{\partial}{\partial\gamma}\text{lev}(\mathcal{J}_{pen}) = \lambda\gamma p(p-1)\left(\frac{1-\gamma}{1-\gamma+\lambda}\right)^{p}\frac{2 - 2\gamma + \lambda + \gamma p}{(1 - \gamma + \gamma p + \lambda)^2(1-\gamma)^2},$$

which is positive for $\gamma > 0$. We also have

$$\frac{1}{p}\frac{\partial}{\partial\lambda}\text{lev}(\mathcal{J}_{pen})\bigg|_{\lambda=0} = \frac{1 + \gamma(p-2)}{(1 - \gamma + \gamma p)(1-\gamma)} \to (1-\gamma)^{-1} \text{ as } p \to \infty,$$

which shows that values of λ close to 0, $\text{lev}(\mathcal{J}_{pen})$ will be much greater when γ is close to 1 than it is for γ close to 0.

In situations where $n \gg p$ with both n and p large, we may be interested in controlling $\text{lev}(\mathcal{J}_{pen})$; for example, we may want to set λ so that $\text{lev}(\mathcal{J}_{pen})$ is neither too close to 0 nor

1. We will (as before) assume that $\mathcal{X}$ is normalized so that the diagonal elements of $\mathcal{X}^T\mathcal{X}$ are all 1. The discussion above suggests that we need to take λ much smaller than the smallest eigenvalue of $\mathcal{X}^T\mathcal{X}$ in order to avoid having $\text{lev}(\mathcal{J}_{pen})$ too close to 1. Assume that $\lambda < \min_{1\le j\le p}\mu_j$; then

$$\sum_{j=1}^{p} \ln\left(\frac{\mu_j}{\mu_j+\lambda}\right) = -\left(\lambda\sum_{j=1}^{p}\mu_j^{-1}\right)(1+r_{n,p})$$

where the remainder term $r_{n,p}$ has the following bound:

$$|r_{n,p}| \le \frac{\lambda}{2}\max_{1\le j\le p}\mu_j^{-1}.$$

Under these assumptions, it follows that for large n and p with $n > p$, we have

$$\text{lev}(\mathcal{J}_{pen}) \approx 1 - \exp\left[-\lambda\,\text{tr}\left((\mathcal{X}^T\mathcal{X})^{-1}\right)\right]$$

if $\max_{1\le j\le p}(\lambda/\mu_j)$ is close to 0.

In addition, we can also consider the leverage of each of the terms in the ridge regression penalty. The leverage of the j-th row of $L(=\lambda^{1/2}I)$ is simply the j-th diagonal element of $\lambda(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1}$ and is the probability (defined in terms of the distribution $\mathcal{P}$ in (5)) that an elemental estimate will have $\widehat{\beta}_j = 0$. We will consider the cases where λ is close to 0 and where λ is "large" in the sense that it exceeds the maximum eigenvalue of $\mathcal{X}^T\mathcal{X}$.

For small values of λ, we consider two cases: $\mathcal{X}^T\mathcal{X}$ invertible and non-invertible. When $\mathcal{X}^T\mathcal{X}$ is invertible, we have

$$\lambda(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1} \approx \lambda(\mathcal{X}^T\mathcal{X})^{-1} - \lambda^2(\mathcal{X}^T\mathcal{X})^{-2}$$

for λ close to 0. Note that the diagonal elements of $(\mathcal{X}^T\mathcal{X})^{-1}$ are simply the variance inflation factors (VIFs)(Marquardt, 1970) of the p predictors since (by assumption) $\mathcal{X}^T\mathcal{X}$ is the correlation matrix of the predictors.

When $\mathcal{X}^T\mathcal{X}$ is not invertible, we can consider the limit of the projection matrix H as defined in (3) as $\lambda \to 0$. The resulting estimate is the OLS estimate with minimum L_2 norm; Hastie *et al.* (2019) refer to this as ridgeless least squares regression. For ridge regression, we have

$$\begin{aligned} H = H_\lambda &= \begin{pmatrix} \mathcal{X}(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1}\mathcal{X}^T & \lambda^{1/2}\mathcal{X}(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1} \\ \lambda^{1/2}(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1}\mathcal{X}^T & \lambda(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1} \end{pmatrix} \\ &\to H_0 = \begin{pmatrix} H_{\mathcal{X}} & 0 \\ 0 & \Gamma_0\Gamma_0^T \end{pmatrix} \end{aligned}$$

as $\lambda \to 0$ where $H_{\mathcal{X}}$ is the projection matrix onto the column space of $\mathcal{X}$ and the columns of Γ_0 are orthogonal eigenvectors of $\mathcal{X}^T\mathcal{X}$ corresponding to the 0 eigenvalues. If we write

$$\mathcal{X}^T\mathcal{X} = (\Gamma_+\ \Gamma_0)\begin{pmatrix} \Lambda_+ & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \Gamma_+^T \\ \Gamma_0^T \end{pmatrix} = \Gamma_+\Lambda_+\Gamma_+^T$$

where Λ_+ is a diagonal matrix of the $r = \text{rank}(\mathcal{X})$ positive eigenvalues of $\mathcal{X}^T\mathcal{X}$ and

$$\Gamma_0 = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_p \end{pmatrix} \text{ and } \Gamma_+ = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_p \end{pmatrix}$$

then the limiting leverages of the rows of $L = \lambda^{1/2} I$ (as $\lambda \to 0$) are $\|\boldsymbol{v}_1\|^2, \cdots, \|\boldsymbol{v}_p\|^2$ (with $\|\boldsymbol{v}_j\|^2 = 1 - \|\boldsymbol{u}_j\|^2$). The elements of $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the principal components loadings for predictor j. Suppose that

$$\mathcal{X} = (\mathbf{X}_1 \ \mathcal{X}_2 B)$$

where the columns of $\mathcal{X}_1$ $(n \times r_1)$ and $\mathcal{X}_2$ $(n \times r_2)$ are linearly independent vectors (with $r_1 + r_2 = r$), and B is a non-zero $r_2 \times (p - r_1)$ matrix whose rank is at least r_2. Then $\mathcal{X}\boldsymbol{a} = \mathbf{0}$ implies that $a_1 = \cdots = a_{r_1} = 0$ and so $\|\boldsymbol{v}_1\|^2 = \cdots = \|\boldsymbol{v}_{r_1}\|^2 = 0$.

If λ is larger than the maximum eigenvalue of $\mathcal{X}^T\mathcal{X}$, we have

$$\begin{aligned} \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} &= \left(I + \frac{1}{\lambda}\mathcal{X}^T\mathcal{X}\right)^{-1} \\ &= I - \frac{1}{\lambda}\mathcal{X}^T\mathcal{X} + \frac{1}{\lambda^2}(\mathcal{X}^T\mathcal{X})^2 - \frac{1}{\lambda^3}(\mathcal{X}^T\mathcal{X})^3 + \cdots \end{aligned}$$

This suggests the following "large λ" approximation for the leverage of the j-th row of $L = \lambda^{1/2} I$:

$$1 - \frac{1}{\lambda} + \frac{1}{\lambda^2}\sum_{k=1}^{p} r_{jk}^2$$

where $\{r_{jk}\}$ are the elements of $\mathcal{X}^T\mathcal{X}$, the pairwise correlations between predictors.

8 Non-quadratic penalized least squares

The theory developed in the previous section has been greatly facilitated by the fact that $\widehat{\boldsymbol{\beta}}$ minimizing (1) is linear in $\boldsymbol{y}$ (as defined in (2)). This allows us to write $\widehat{\boldsymbol{\beta}}$ as a weighted average of elemental estimates where (critically) the weights depend only on the rows of matrices $\mathcal{X}$ and L, and not on the vector $\boldsymbol{y}$.

In practice, non-quadratic penalties, such as the LASSO (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), and SCAD (Fan and Li, 2001), are commonly used; these methods can produce exact 0 estimates of parameters and therefore are useful for model selection. Extending the notion of leverage to these methods is somewhat problematic as the resulting estimates are non-linear in $\boldsymbol{y}$ and local linear approximations are often inadequate as the penalty terms often contain non-differentiable components. Tibshirani and Taylor (2012) are able to resolve these problems in defining equivalent degrees of freedom for the LASSO.

As an example, consider the LASSO. We will assume the same conditions about the matrix $\mathcal{X}$ as in the case of ridge regression: The intercept is estimated by the average of $y_1, \cdots, y_n$ and the predictors are centred and scaled so that $\mathcal{X}^T\mathcal{X}$ is the correlation matrix of the predictors. The LASSO estimate $\widehat{\boldsymbol{\beta}}_\lambda$ then minimizes

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where $\|\boldsymbol{\beta}\|_1$ is the L_1-norm of $\boldsymbol{\beta}$.

Suppose that λ is sufficiently small that all the components of $\widehat{\boldsymbol{\beta}}_\lambda$ are non-zero. Then we can write $\widehat{\boldsymbol{\beta}}_\lambda$ in a pseudo-linear form

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_\lambda &= \left(\mathcal{X}^T\mathcal{X} + \frac{\lambda}{2}D^{-1}(\widehat{\boldsymbol{\beta}}_\lambda)\right)^{-1}\mathcal{X}^T\boldsymbol{y}_0 \\ &= \left(\mathcal{X}^T\mathcal{X} + \frac{\lambda}{2}D^{-1}(\widehat{\boldsymbol{\beta}}_\lambda)\right)^{-1}\begin{pmatrix}\mathcal{X} \\ (\lambda/2)^{1/2}D^{-1/2}(\widehat{\boldsymbol{\beta}}_\lambda)\end{pmatrix}^T\boldsymbol{y} \qquad (9)\end{aligned}$$

where $\boldsymbol{y}_0 = (y_1, \cdots, y_n)^T$ and $D(\widehat{\boldsymbol{\beta}}_\lambda)$ is a diagonal matrix with elements $|\widehat{\beta}_1|, \cdots, |\widehat{\beta}_p|$. In this case, we could apply the results of the previous section with the matrix $\lambda D^{-1}(\widehat{\boldsymbol{\beta}}_\lambda)/2$ (which depends on $\boldsymbol{y}_0$) playing the role of A in our results.

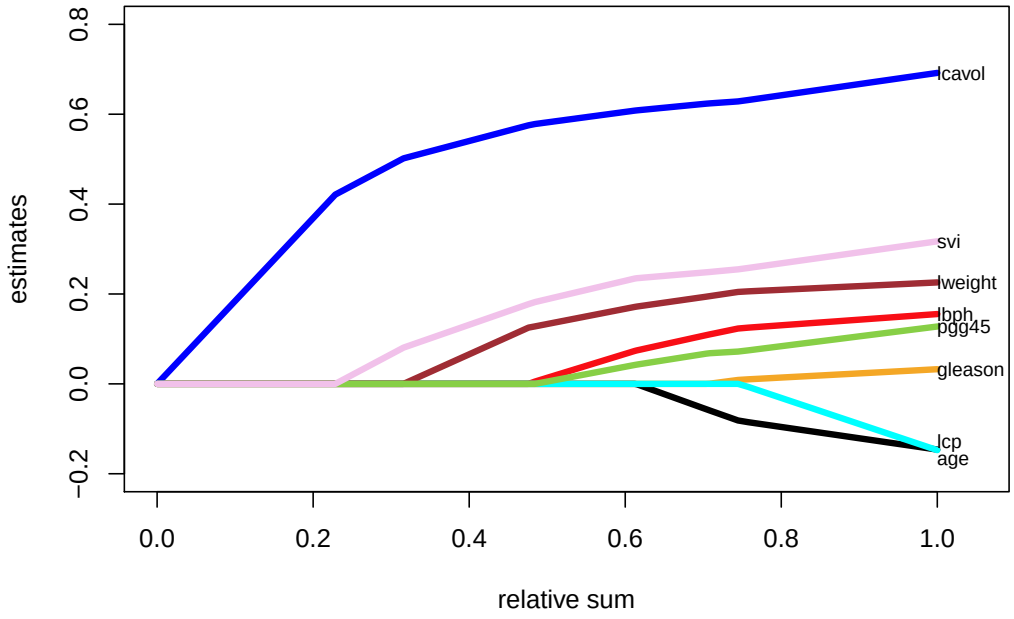

Figure 6: LASSO plot: Plot of the 8 estimates as a function of $\|\boldsymbol{\beta}_\lambda\|_1/\|\boldsymbol{\beta}_0\|_1$.

The situation seemingly becomes somewhat more complicated when λ is sufficiently large so that one or more of the components of $\widehat{\boldsymbol{\beta}}_\lambda$ equals 0. However, note that we can write the

projection matrix arising in (9) as

$$\begin{pmatrix} \mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda) \\ (\lambda/2)^{1/2}I \end{pmatrix} \left((\mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda))^T (\mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda)) + \frac{\lambda}{2} I \right)^{-1} \begin{pmatrix} \mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda) \\ (\lambda/2)^{1/2}I \end{pmatrix}^T \tag{10}$$

If the j component of $\widehat{\boldsymbol{\beta}}_\lambda$ equals 0 then the j column of $\mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda)$ will be a vector of 0s and the rank of $\mathcal{X}D^{1/2}(\widehat{\boldsymbol{\beta}}_\lambda)$ will be reduced (relative to the rank of $\mathcal{X}$) by the number of 0 components in $\widehat{\boldsymbol{\beta}}_\lambda$. For a given value of λ, we can apply the results given earlier to the projection matrix in (10). In particular, if the j component of $\widehat{\boldsymbol{\beta}}_\lambda$ equals 0 then the $n+j$ diagonal element of the projection matrix in (10) equals 1.

As an illustration, we consider the prostate cancer data described in Tibshirani (1996), where the relationship between the logarithm of prostate specific antigen (PSA) and 8 predictors (prognostic variables). Figure 6 shows that LASSO plot for these data while Figures 7 and 8 show the corresponding influence of the "zero constraint" for each of the 8 predictors for LASSO and ridge regression, respectively. The predictors here are not particular collinear – the largest VIF is approximately 3.1, which is well below the usual thresholds for high collinearity. As a result, the influence of the penalty term for each predictor (that is, the "zero constraint") varies little across predictors as λ increases. Not surprisingly, the influence of the penalty terms across predictors for the LASSO has a greater variation due to the nature of the LASSO penalty.

9 Final comments

In this paper, we have defined leverage as a probability arising from the probability measure $\mathcal{P}$ defined in (5) whose marginal distributions depend on the projection matrix H in (3). Berman (1988) extends Jacobi's result to idempotent matrices of the form

$$H = X(Z^T X)^{-1} Z^T,$$

which arise, for example, in generalized least squares (where $Z = \Omega X$ for some $n \times n$ matrix Ω) and instrumental variables regression (where Z is a matrix of instrumental variables). In this case,

$$\widehat{\boldsymbol{\beta}} = (X^T Z)^{-1} Z^T \boldsymbol{y} = \sum_s \mathcal{Q}(s) \widehat{\boldsymbol{\beta}}_s$$

where now

$$\mathcal{Q}(s) = \left| X_s (X^T Z)^{-1} Z_s^T \right| \quad \text{and} \quad \sum_s \mathcal{Q}(s) = 1.$$

However, $\mathcal{Q}(s)$ is not necessarily a probability measure in the idempotent case: We may have $\mathcal{Q}(s) < 0$ or $\mathcal{Q}(s) > 1$ for some s. In this latter case, Propositions 1 and 2 still hold in a mathematical sense, replacing the probabilities with sums of $\mathcal{Q}(s)$ over a collection of s.

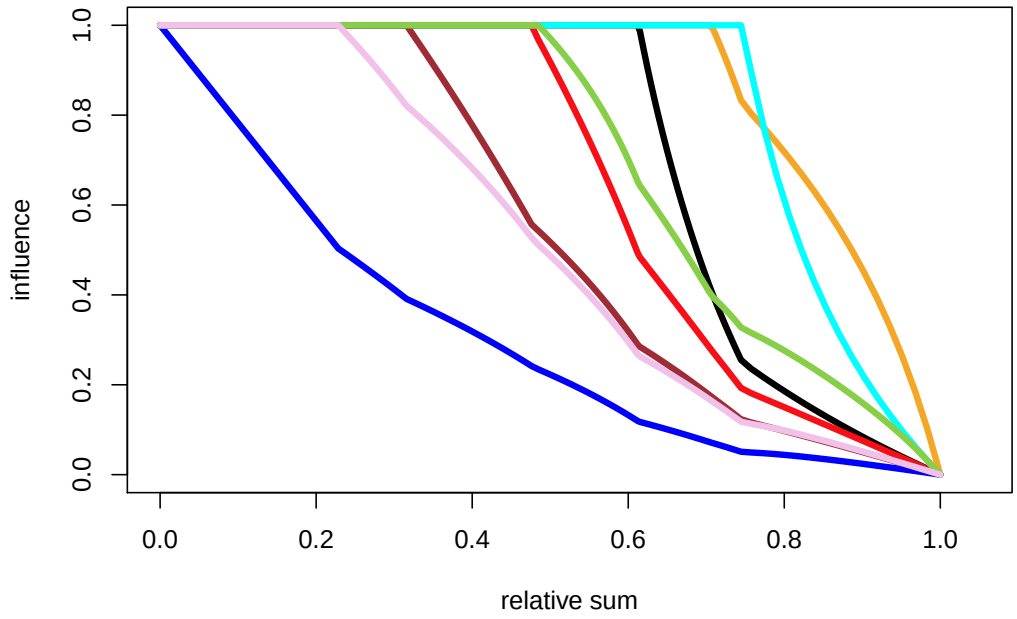

Figure 7: Influence plot for LASSO

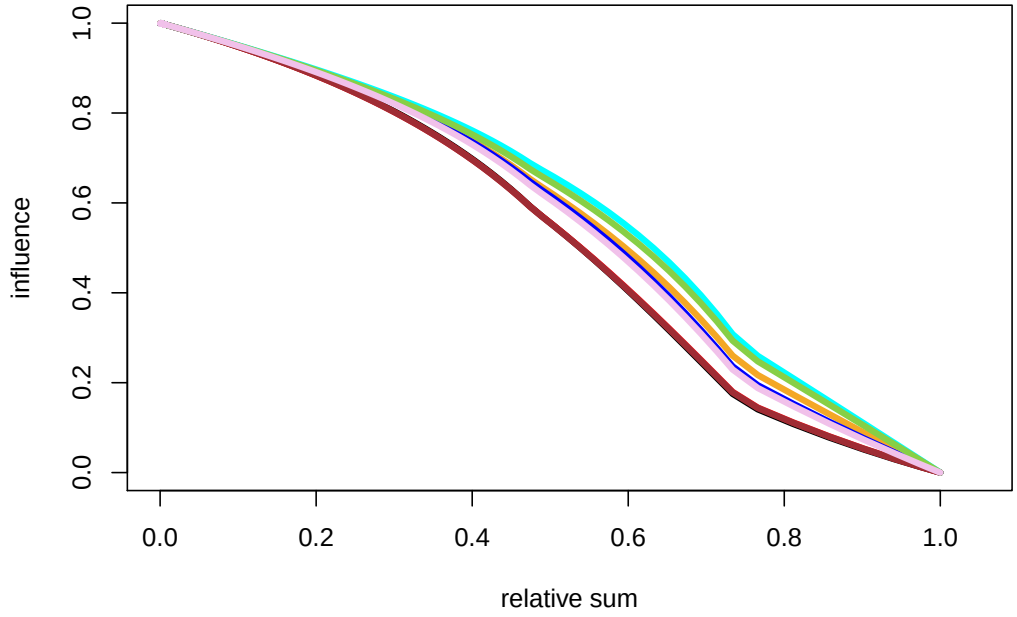

Figure 8: Influence plot for ridge regression

An important limitation of this paper is that the approach to defining leverage applies to a relatively small class of penalized least squares estimates, namely those where the penalty term is itself quadratic. In section 8, we outlined how we might extend the definition to non-quadratic penalties; extensions to non-quadratic objective functions (for example, penalized maximum likelihood) are less obvious although the basic idea would be to approximate the objective function near its minimizing value by a quadratic function.

Appendix: Proof of Proposition 1

Define the $k \times k$ matrix

$$H_{i_1 \cdots i_k}(\boldsymbol{t}) = \exp(t_{i_1} + \cdots + t_{i_k}) \begin{pmatrix} \boldsymbol{x}_{i_1}^T \\ \vdots \\ \boldsymbol{x}_{i_k}^T \end{pmatrix} \left(\sum_{i=1}^{n+r} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} (\boldsymbol{x}_{i_1} \cdots \boldsymbol{x}_{i_k})$$

and define for $1 \leq i, j \leq n + r$,

$$h_{ij}(\boldsymbol{t}) = \boldsymbol{x}_j^T \left(\sum_{i=1}^{n+r} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i.$$

It suffices to show that

$$\frac{\partial^k}{\partial t_{i_1} \cdots \partial t_{i_k}} \varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t}) \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right|. \tag{11}$$

We will prove (11) by induction using Jacobi's formula (Golberg, 1972)

$$\frac{d}{dt}|K(t)| = \mathrm{tr}\left(\mathrm{adj}(K(t)) \frac{d}{dt} K(t) \right) = |K(t)| \mathrm{tr}\left(K^{-1}(t) \frac{d}{dt} K(t) \right)$$

where $\mathrm{adj}(K(t))$ is adjugate (the transpose of the cofactor matrix) of $K(t)$ as well as the identity

$$\left| \begin{pmatrix} D & \boldsymbol{v} \\ \boldsymbol{v}^T & a \end{pmatrix} \right| = a|D| - \boldsymbol{v}^T \mathrm{adj}(D) \boldsymbol{v} \tag{12}$$

where a is a real number, $\boldsymbol{v}$ a vector of length k, and D a $k \times k$ matrix. For $k = 1$, we have

$$\begin{aligned} \frac{\partial}{\partial t_{i_1}} \varphi(\boldsymbol{t}) &= \varphi(\boldsymbol{t}) \mathrm{tr} \left\{ \left(\sum_{i=1}^{n+r} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \exp(t_{i_1}) \boldsymbol{x}_{i_1} \boldsymbol{x}_{i_1}^T \right\} \\ &= \varphi(\boldsymbol{t}) \exp(t_{i_1}) \boldsymbol{x}_{i_1}^T \left(\sum_{i=1}^{n+r} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_{i_1} \\ &= \varphi(\boldsymbol{t}) \exp(t_{i_1}) h_{i_1 i_1}(\boldsymbol{t}) \\ &= \varphi(\boldsymbol{t}) \left| H_{i_1}(\boldsymbol{t}) \right|. \end{aligned}$$

Now suppose that (11) holds for some $k < p$ and set $\ell = k + 1$. Then

$$\begin{aligned} \frac{\partial^\ell}{\partial t_{i_1} \cdots \partial t_{i_\ell}} \varphi(\boldsymbol{t}) &= \frac{\partial}{\partial t_{i_\ell}} \left\{ \varphi(\boldsymbol{t}) \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right| \right\} \\ &= \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right| \frac{\partial}{\partial t_{i_\ell}} \varphi(\boldsymbol{t}) + \varphi(\boldsymbol{t}) \frac{\partial}{\partial t_{i_\ell}} \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right|. \end{aligned}$$

First,

$$\frac{\partial}{\partial t_{i_\ell}}\varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t})\,|H_{i_\ell}(\boldsymbol{t})| = \varphi(\boldsymbol{t})\exp(t_{i_\ell})h_{i_\ell i_\ell}(\boldsymbol{t}).$$

Second,

$$\varphi(\boldsymbol{t})\frac{\partial}{\partial t_{i_\ell}}\,|H_{i_1\cdots i_k}(\boldsymbol{t})| \;\; = \;\; \varphi(\boldsymbol{t})\left\{\mathrm{tr}\left(\mathrm{adj}(H_{i_1\cdots i_k}(\boldsymbol{t}))\frac{\partial}{\partial t_{i_\ell}}H_{i_1\cdots i_k}(\boldsymbol{t})\right)\right\}$$

with

$$\frac{\partial}{\partial t_{i_\ell}}H_{i_1\cdots i_k}(\boldsymbol{t}) = -\exp(t_{i_1}+\cdots+t_{i_k}+t_{i_\ell})\begin{pmatrix} h_{i_1 i_\ell}(\boldsymbol{t}) \\ \vdots \\ h_{i_k i_\ell}(\boldsymbol{t}) \end{pmatrix}(h_{i_1 i_\ell}(\boldsymbol{t})\cdots h_{i_k i_\ell}(\boldsymbol{t}))\,.$$

Applying (12) with

$$a = h_{i_\ell i_\ell}(\boldsymbol{t}), \quad D = H_{i_1\cdots i_k}(\boldsymbol{t}), \quad \text{and} \quad \boldsymbol{v} = \begin{pmatrix} h_{i_1 i_\ell}(\boldsymbol{t}) \\ \vdots \\ h_{i_k i_\ell}(\boldsymbol{t}) \end{pmatrix},$$

we get

$$\frac{\partial^\ell}{\partial t_{i_1}\cdots\partial t_{i_\ell}}\varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t})\,|H_{i_1\cdots i_\ell}(\boldsymbol{t})|$$

and the conclusion follows by setting $\boldsymbol{t} = \boldsymbol{0}$.

References

Barrett, B., Gray, J.B.: Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. Computational statistics and data analysis. **26**, 39–52 (1997)

Berman, M.: A theorem of Jacobi and its generalization. Biometrika, **75**, 779–783 (1988)

Bessenyei, M., Páles, Z.: Higher-order generalizations of Hadamard's inequality. Publicationes Mathematicae Debrecen. **61**. 623–643 (2002)

Breiman, L.: Bagging predictors. Machine learning. **24**. 123–40 (1996)

Bullen, P.S.: Error estimates for some elementary quadrature rules. Publikacije Elektrotehničkog fakulteta. Seijia Matematika i fizika. **602/633**. 97–103. (1978)

Chatterjee, S., Hadi, A.S.: Influential observations, high leverage points, and outliers in linear regression. Statistical Science, **1**, 379–393 (1986)

Choi, M.-D., Wu, P.Y.: Convex combinations of projections. Linear algebra and its applications. **136**, 25–42 (1990)

Clerc Bérod, A., Morgenthaler, S.: A close look at the hat matrix. Student, **2**, 1–12 (1997)

Cornea-Madeira, A.: The explicit formula for the Hodrick-Prescott filter in a finite sample. Review of economics and statistics, **99**, 314–318 (2017)

Cook, R.D., Weisberg, S.: Residuals and Influence in Regression (Chapman and Hall, London) (1982)

Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numerische mathematik **31**, 377–403 (1978)

Derezinski, M., Warmuth, M.K., Hsu, D.J.: Leveraged volume sampling for linear regression. Advances in Neural Information Processing Systems. 2505–2514 (2018)

Dobriban, E., Wager, S.: High-dimensional asymptotics of prediction: ridge regression and classification. Annals of Statistics. **46**, 247–279 (2018)

Draper, N.R., John, J.A.: Influential observations and outliers in regression. Technometrics, **23**, 21–26 (1981)

Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.: Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research. **1**3, 3475–3506 (2012)

Drineas, P., Mahoney, M.W., Muthukrishnan, S., Sarlós, T.: Faster least squares approximation. Numerische Mathematik. **117**, 219–249 (2011)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. **96**, 1348–1360 (2001)

Gao, K.: Statistical inference for algorithmic leveraging. arXiv preprint arXiv:1606.01473 (2016)

Golberg, M.A.: The derivative of a determinant. The American Mathematical Monthly. **79**, 1124–1126 (1972)

Hamilton, J.D.: Why you should never use the Hodrick-Prescott filter. Review of Economics and Statistics. **100**, 831–843 (2018)

Hastie, T., Montanari, A., Rosset, S., Tibshirani, R.: Surprises in high-dimensional ridgeless least squares interpolation. arXiv: 1903.08560 (2019)

Hastie, T., Tibshirani, R.: Generalized Additive Models (Chapman and Hall, London) (1990)

Hellton, K.H., Lingjaerd,C., De Bin, R. Influence of single observations on the choice of the penalty parameter in ridge regression. ArXiv preprint arXiv: 1911.03662 (2019)

Hoaglin, D.C., Welsch, R.E.: The hat matrix in regression and ANOVA. The American Statistician, **32**, 17–22 (1978)

Hodrick, R.J., Prescott,E.C.: Postwar US business cycles: an empirical investigation. Journal of Money, Credit, and Banking. **29**, 1–16 (1997)

Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, **12**, 55–67 (1970)

Hutchinson, M.F.: A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Communications in Statistics-Simulation and Computation, **19**, 433–450 (1990)

Jacobi, C.G.J.: De formatione et proprietatibus Determinatium. Journal für die reine und angewandte Mathematik, **22**, 285–318 (1841)

Knight, K.: Elemental estimates, influence, and algorithmic leveraging. In: Nonparametric Statistics, 3rd ISNPS Avignon. 219–231 (2019)

Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research. **16**, 861–911 (2015)

Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. **12**, 591–612 (1970)

Mayo, M.S., Gray, J.B.: Elemental subsets: the building blocks of regression. The American Statistician. **51**, 122–129 (1997)

Nurunnabi, A.A.M., Hadi, A.S.,Imon, A.H.M.R.: Procedures for the identification of multiple influential observations in linear regression. Journal of Applied Statistics. **41**, 1315–1331 (2014)

Skilling, J.: The eigenvalues of mega-dimensional matrices. In Maximum Entropy and Bayesian Methods. 455–466 (1989)

Subrahmanyam, M.: A property of simple least squares estimates. Sankhya, Series B. **34**, 355-356 (1972)

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B). **58**, 267–288 (1996)

Tibshirani, R., Taylor, J.: Degrees of freedom in lasso problems. Annals of Statistics. **40**, 1198–1232 (2012)

Whittaker, E.T.: On a new method of graduation. Proceedings of the Edinburgh Mathematical Society. **41**, 63–75 (1922)

Wolpert, D.H.: Stacked generalization. Neural networks **5**, 241–259 (1992)

Zou, H, Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society (Series B). **67**, 301–320 (2005)